Tracking Human Migration from Online Attention

Carmen Vaca-Ruiz[1,2(✉)], Daniele Quercia[2], Luca Maria Aiello[2], and Piero Fraternali[1]

[1] Politecnico di Milano, Milan, Italy
{vacaruiz,fraterna}@elet.polimi.it
[2] Yahoo Research, Barcelona, Spain
{dquercia,alucca}@yahoo-inc.com

Abstract. The dynamics behind human migrations are very complex. Economists have intensely studied them because of their importance for the global economy. However, tracking migration is costly, and available data tends to be outdated. Online data can be used to extract proxies for migration flows, and these proxies would not be meant to replicate traditional measurements but are meant to complement them. We analyze a random sample of a microblogging service popular in Brazil (more than 13M posts and 22M reposts) and accurately predict the total number of migrants in 35 Brazilian cities. These results are so accurate that they have promising implications in monitoring emerging economies.

1 Introduction

For census agencies, migrations are difficult to track in the developed countries, let alone in developing ones. In emerging economies, authorities rely on inaccurate, outdated, de-contextualized census data even for the local population [15].

Migrants who have left their home country searching for better opportunities rely also on electronic communication to maintain their bonds with their home communities [3]. Publishing and 'consuming' content such as news and photos in online platforms is "a parcel of everyday life in transnational families" [2]. Previous studies have found that indicators characterizing offline communities (e.g., economic deprivation) can be extracted from online data (e.g., use of emotion words in Twitter) [22]. Therefore, we propose to consider online data in Brazil and track the number of migrants in a city by considering the interaction between users who live in the city and those outside.

Our main contribution is to propose a set of metrics extracted from online data to estimate migration levels. These metrics reflect the intuition that the higher the number of migrants in a city, the more online interactions between users in the city and those outside it. We compute these metrics for 35 cities in a Yahoo Meme dataset that includes more than 13M posts and 22M reposts exchanged between users in more than 1K cities around the world. We find that the proposed metrics work, in that, they correlate with the number migrants reported by the Brazilian census authority. By then combining these metrics in a linear regression model, we show that the model fits the data extremely well (the *Adj.* $R^2 = 0.61$).

J. Nin and D. Villatoro (Eds.): CitiSens 2013, LNAI 8313, pp. 73–83, 2014.
DOI: 10.1007/978-3-319-04178-0_7,

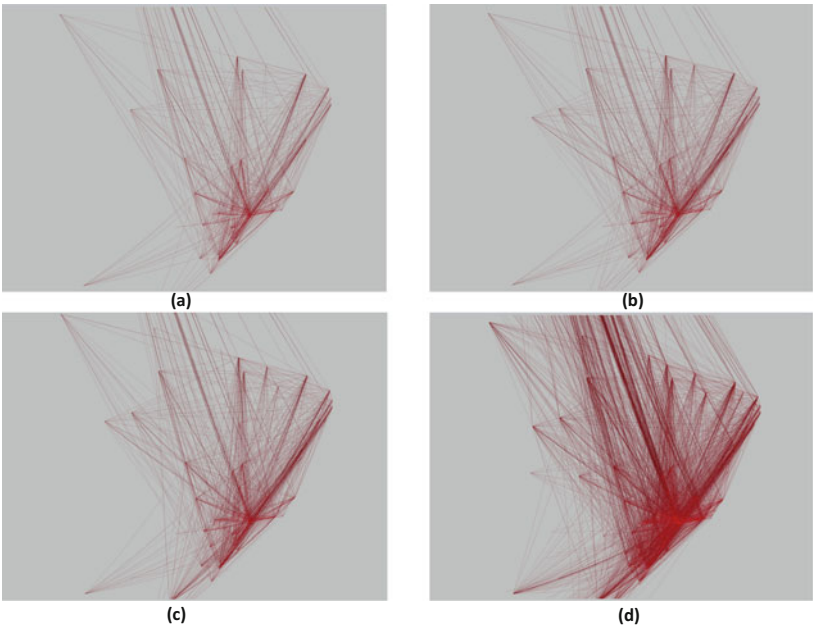

Fig. 1. Evolution of the follower graph: edges connect the geographical locations of the users in our dataset. The picture shows the cumulative set of graph edges after the (a) 1st, (b) 2nd, (c) 3rd and (d) 7th month after the platform launch. The brightest point in (a) corresponds to the city of Sao Paulo.

2 Dataset

Yahoo Meme was a microblogging platform, similar to Twitter, with the exception that users can post content of any length or type (text, pictures, audio, video), being text and pictures the more frequently posted content. In addition to posting, users could also *follow* other users, *repost* others' content, and *comment* on it. In this study, we use a random sample of interactions on Yahoo Meme from its birth in 2009 until the day it was discontinued in 2012 (Table 1). Despite its moderate popularity in USA, Yahoo Meme was popular in Brazil, as witnessed by the fact that the top 45 cities in terms of number of interactions are all located there. Reposting was the main activity in the service (22M sample records) compared to comments (4M). We extract the users who posted the content in our sample and georeference them based on their IP addresses using a Yahoo service. We remove the users for whom we did not obtain results at city level (e.g., users employing proxy servers to connect to the Internet) obtaining 80 K users. For this set of users and their respective posts, we extract all the repost *cascades* and the follower relationships. Month after month, users across different Brazilian cities tended to intensify their follower connections till reaching a certain stability at month 7 after the platform launch (Fig. 1).

Table 1. Yahoo Meme dataset statistics

Property	Value
Number of users	80 K
Number of posts	13M+
Number of reposts	22M+
Number of follower links	19M+
Number of comments	4M
Number of reposts cascades	1.4M
Number of cities	1.3 K
Number repost edges between cities	25 K

To attain geographic representability, we ascertain that the number of users in the top Brazilian cities in our dataset is significantly correlated with the number of Internet users (Fig. 2). As a result, any city outside the confidence area calculated (outlier) is excluded from the study. This leaves us with 35 cities, and we will see that such a number grants statistical significant results. That is because we are left with 1.4M repost cascades whose original content was produced in the 35 cities and was consumed across the world.

3 Attention Metrics

It has been shown that migrants maintain their strong ties in their home countries mainly using digital means [2]. We thus expect that studying online interactions in Yahoo! Meme across geographic areas would result in good estimators of migration flows. More specifically, we connect places every time that a user u_i located in city i interacts with a user located in city j either by reposting u_i's content or by following him/her. The volume of such connections is then correlated with migration rates for 35 cities in Brazil. We consider migrants from Brazil itself and from the rest of the world.

Previous studies have shown that interactions on social media cannot be quantified with simple metrics such as popularity or number of followers but they are best characterized with metrics that also reflect the extent to which content is re-shared or liked [1,6,23,30,32]. That is because social media users make specific decisions about the content they want to consume or who they wish to follow. Such decisions are taken based on offline social ties [31], homophily, and physical distance [25].

We thus resort to attention metrics, and these metrics capture the attention that a city's users are able to attract from the *Rest of the World* and from other *Brazilian* cities:

Cross Border Attention. Our first set of attention metrics for city i is defined as the number of reposts that the city has attracted from the rest of the world (ROW_i^{repost}) or from other Brazilian cities (BR_i^{repost}), normalized with respect to the total number n_i of users in that city:

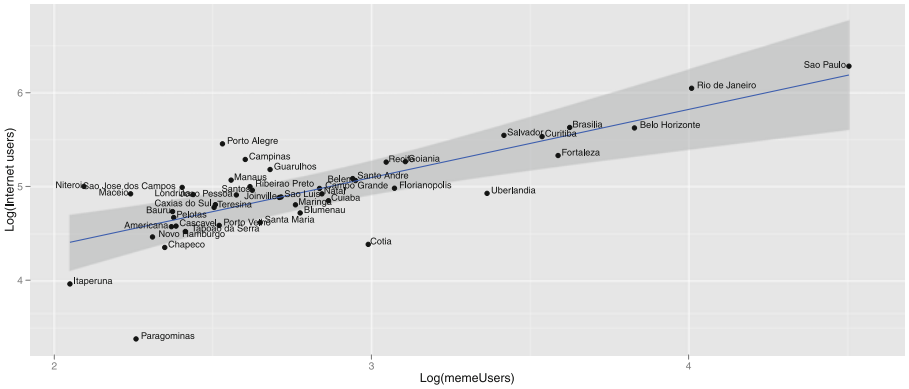

Fig. 2. Number of users in our sample versus number of Internet users. Both quantities are log-transformed. Regression line and 95 % confidence intervals are shown

$$ROW_i^{repost} = \frac{out_i}{n_i}, BR_i^{repost} = \frac{out'_i}{n_i}$$

where out_i is the number of times a post originated in city i has been reposted outside it (the world excluding Brazil); out'_i, instead, counts the reposts received outside the city but inside Brazil.

We repeat the same definition considering now the number of cross-borders followers attracted by users in city i:

$$ROW_i^{followers} = \frac{outf_i}{n_i}, BR_i^{followers} = \frac{outf'_i}{n_i}$$

where $outf_i$ is the number of times a user in city i has been followed by a user outside it (the world excluding Brazil); $outf'_i$, instead, counts the follower links outside the city but inside Brazil. As a result, we obtain the first four metrics.

Authority. The previous metrics consider all cities equally. However, certain cities might be more central to migration flows than others. To capture this concept of centrality, we built an *attention graph* using reposts. This is a weighted directed graph where nodes are cities, and directed weighted edges (i, j, w) represent the volume w of reposts between city j where the *reposter* lives, and city i where the original *poster* lives. Self-edges are allowed as many reposts occur between users living in the same city. The resulting *attention graph* has 1,310 nodes and 25 K weighted edges (Fig. 3). Then, we measure the 'authority' index of each city using the HITS algorithm [14]. In the HITS algorithm the autorithy centrality of a vertex is defined to be proportional to the aggregated values of the hub centrality indexes that point to it. For a city i, the two indexes as defined as follows:

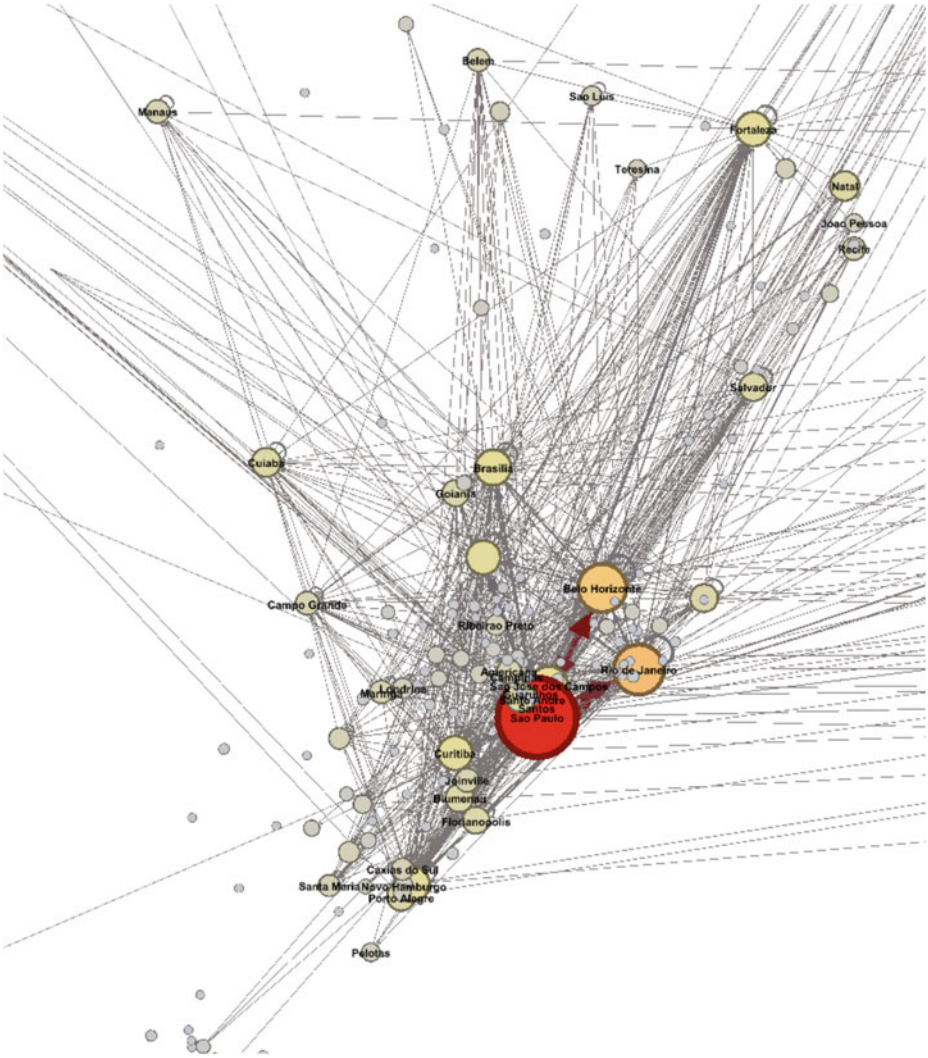

Fig. 3. Attention graph whose nodes are cities and whose weighted edges reflect the intensity of reposting between cities' users. Size and color of the nodes are proportional to the node degree. The network was plotted using the GeoLayout plugin of the Gephi software package [14].

$$Authority_i = \alpha \cdot \sum_{j \in C} A_{ij} Hub_j,$$

$$Hub_i = \beta \cdot \sum_{j \in C} A_{ji} Authority_i,$$

where α and β are constants, C is the set of cities in our dataset and A is the *attention graph*'s corresponding city adjacency matrix.

The Authority index calculated by the HITS algorithm is more informative for the vertex centrality in directed networks than simpler measures such as the number of incident edges or indegree centrality [12] and, thus, it better captures the importance of a node in the network.

We calculate the correlation among each pair of the five metrics: ROW_i^{repost}, BR_i^{repost}, $ROW_i^{followers}$, $BR_i^{followers}$, $Authority_i$ (Fig. 4) and observe that they are all correlated with each other. That is why, when we will run our predictions, we will account for interaction effects.

4 Correlations Between Attention and Migration

From the 2010 data provided by the Brazilian census bureau[1], we compute two *migration rates* for each of the 35 cities: m_{ROW} is the number of people coming from other countries and m_{BR} is that from other Brazilian cities. Both values are normalized by city population. We then correlate these two migration rates with our five attention metrics. To account for skewness, the metrics are log-transformed. The results obtained are statistically significant, with at least p-value < 0.05.

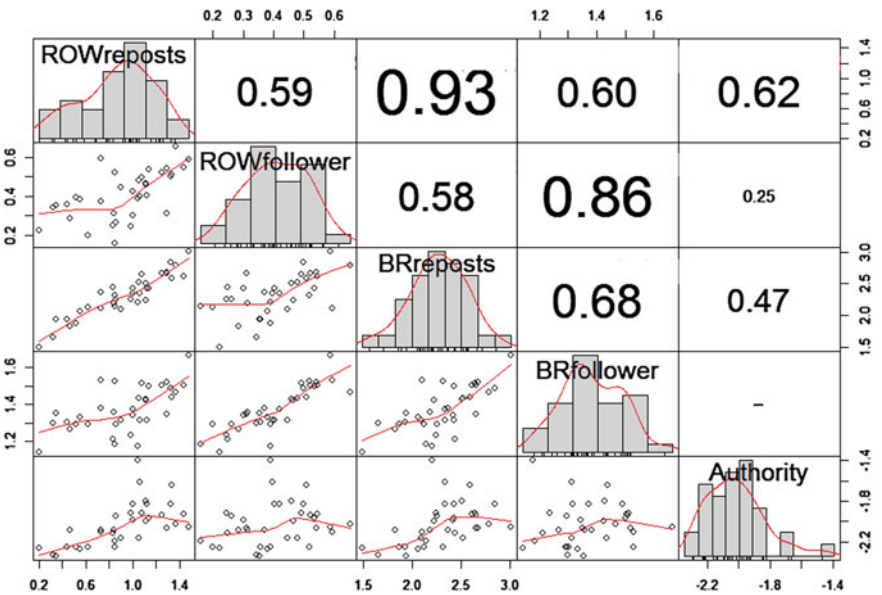

Fig. 4. Correlations among the five attention metrics. We observe that the *ROW* attention metrics are correlated among each other more than they are with the Authority metric. Values are log-transformed.

[1] http://www.ibge.gov.br

Reposts and Follower metrics. We find positive correlations between migration rates and attention received by the rest of the world: $r = 0.28$ for *attention* computed on reposts, and $r = 0.33$ for attention computed on number of followers. Stronger correlations are also found for attention received from other Brazilian cities: $r = 0.33$ for attention computed on reposts, and $r = 0.46$ for attention computed on number of followers.

Authority metric. Since the authority measure can be only computed on the aggregate (Brazil plus rest-of-the-world) dataset, we should correlate the authority measure with the *total* number of migrants ($m_{ROW} + m_{BR}$). In so doing, we obtain, again, a positive correlation $r = 0.32$.

5 Predicting Migration from Attention

We model the number of migrants as a linear combination of the five attention metrics. This is what we call Model1:

$$\begin{aligned} log(MigrantsNumber_i) = \alpha + \beta_1 \cdot log(ROW_i^{repost}) + \\ \beta_2 \cdot log(ROW_i^{followers}) + \beta_3 \cdot log(BR_i^{repost}) + \\ \beta_4 \cdot log(BR_i^{followers}) + \beta_5 \cdot log(Authority_i) + \\ \epsilon_i \end{aligned} \tag{1}$$

We also build a model to account for the pairwise interactions effects between indicators:

$$\begin{aligned} log(MigrantsNumber_i) = \alpha + \beta_1 \cdot log(ROW_i^{repost}) + \\ \beta_2 \cdot log(ROW_i^{followers}) + \beta_3 \cdot log(BR_i^{repost}) + \\ \beta_4 \cdot log(BR_i^{followers}) + \beta_5 \cdot log(Authority_i) + \\ \gamma_m \cdot Interactions_{im} + \epsilon_i \end{aligned} \tag{2}$$

where $Interactions_{im}$ accounts for the pairwise interactions among the five attention metrics. This is model 2 (Table 2).

To account for Internet penetration rates and population, we build a model adding these two census variables

$$\begin{aligned} log(MigrantsNumber_i) = \alpha + \beta_1 \cdot log(ROW_i^{repost}) + \\ \beta_2 \cdot log(ROW_i^{followers}) + \beta_3 \cdot log(BR_i^{repost}) + \\ \beta_4 \cdot log(BR_i^{followers}) + \beta_5 \cdot log(Authority_i) + \\ + \mu_i Internet_i + \rho_i Population_i + \\ + \gamma_m Interactions_{im} + \epsilon_i \end{aligned} \tag{3}$$

where $Internet_i$ is the city's Internet's penetration rate, $Population_i$ is the city's population, and ϵ_i is the error term. This is Model 3. We control for Internet

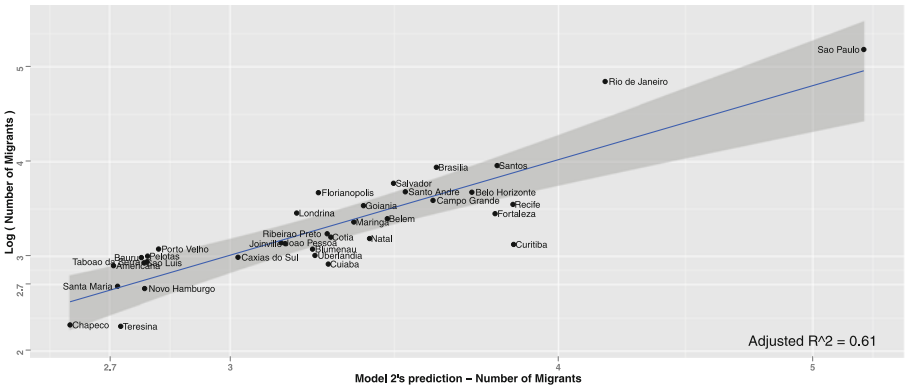

Fig. 5. Predicted values versus actual values calcuated by Model 2 (*Adj.* R^2=0.61) that includes the five attention metrics and their pairwise interactions. The model's prediction error is low: its Mean Absolute Error is 0.21.

Table 2. *Adj.* R^2 for different models predicting city i's number of migrants. Model 1's predictors are the five attention metrics $Attention_{im}$, Model 2 adds their interaction effects, Model 3 controls for the city's Internet penetration rates and population. All *p-values* are < 0.001.

Model	Predictors	$Adjusted.R^2$
1	$\{Attention_{im}\}$	**0.54**
2	$\{Attention_{im}\} + \{Interactions_{im}\}$	**0.61**
3	$\{Attention_{im}\} + \{Interactions_{im}\} + Internet_i + Population_i$	**0.70**

penetration because it is associated with online activity, and for city size because larger cities tend to be economically prosperous and enjoy "increasing returns to scale": a city becomes more attractive as it grows [12].

By computing the beta coefficients of model 2, the one with the best performance (without census data), we find that *cross border attention* in terms of followers accounts for 22 % of the model's explanatory power, while the *cross border attention* for reposts explains 18 %. *Authority* attention, instead, only explains 7 % of the variance. As for model 2's accuracy, the model achieves a Mean Absolute Error (MAE) of 0.21 on a logarithmic scale, where the minimum value is 2.6 and maximum is 5.23, meaning that, on average, the model predicts the log of the number of migrants within 1.16 % of its true value. Figure 5 plots the values predicted by model 2 against actual ones. Rio de Janeiro, one of the most international Brazilian cities, is one outlier for which the number of migrants level is higher than the predicted value.

6 Related Work

Real-life Processes and Social Media. Email exchanges have been used to track migration flows among developed and developing countries [26]. Also, Quercia *et al.* have shown a correlation between the sentiment expressed in tweets originated by residents of London neighborhoods and the neighborhoods' well-being [22].

In the last few years, there have appeared some initiatives for measuring socio-economic conditions of city residents in developing countries using online data. For example, the United Nations and the World Bank have recently launched a program called "Data4Good". This promotes the use of (currently untapped) digital data for, say, improving poverty measurement ("How can we measure poverty more often and more accurately?") or dealing with corruption in international investment projects ("Can we detect fraud by looking at aid data?"). Recently, Orange released an anonymized dataset of mobile phone calls in Côte d'Ivoire, and launched a challenge in which researchers had to predict economic indicators from the activity metrics extracted from the call records [17]. Our research complements this line of work by proposing a set of metrics that can be applied to data extracted from any data source that reflects social exchanges, including social media data.

Migration. Davis *et al.* [8] conducted a study of human mobility using data published by the World Bank. They built a network of countries based on migration flows, and found that the most well connected countries remain stable over time and that migration is directed towards low and mid degree countries.

7 Conclusion

We have shown that online metrics are effective at predicting number of migrants. These metrics are particularly useful in developing countries, where economic changes happen at fast pace. As part of future work, we will study socio-economic indicators other than migration rates, and we will start with GDP and social capital.

Acknowledgments. Carmen Vaca Ruiz's research work has been funded by SENESCYT and ESPOL, Ecuador.

References

1. Asur, S., Huberman, B.A., Szabo, G., Wang, C.: Trends in social media: persistence and decay. In: Proceedings of the 5th AAAI Conference on Weblogs and Social Media (ICWSM) (2011)
2. Baerenholdt, J.O., Granås, B.: Mobility and Place: Enacting Northern European Peripheries. Ashgate Publishing Ltd., Hardcover (2008)
3. Bates, J., Komito, L.: Migration, community and social media. Transnationalism in the Global City, vol. 6. University of Deusto, Bilbao (2012)

4. Boucher, G., Grindsted, A., Vicente, T.L. (eds.): Transnationalism in the Global City. Universidad de Deusto, Bilbao (2012)
5. Brodersen, A., Scellato, S., Wattenhofer, M.: Youtube around the world: geographic popularity of videos. In: Proceedings of the 21st ACM Conference on World Wide Web (WWW) (2012)
6. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.: Measuring user influence in twitter: the million follower fallacy. In: Proceedings of the 4th AAAI Conference on Weblogs and Social Media (ICWSM) (2010)
7. Datta, A.: Human Migration: A Social Phenomenon. Mittal Publications, New Delhi (2003)
8. Davis, K.F., D'Odorico, P., Laio, F., Ridolfi, L.: A complex network perspective. PloS One **8**(1), e53723 (2013)
9. Eagle, N., Macy, M., Claxton, R.: Network diversity and economic development. Science **328**(5981), 1029–1031 (2010)
10. Favell, A., Feldblum, M., Smith, M.P.: The human face of global obility: a research agenda. Society **44**(2), 15–25 (2007)
11. Ghosh, R., Lerman, K.: Predicting influential users in online social networks. In: Proceedings of the 4th AAAI Conference on Weblogs and Social Media (ICWSM) (2010)
12. Glaeser, E.L., Kohlhase, J.E.: Cities, regions and the decline of transport costs. Reg., Sci. **83**(1), 197–228 (2004)
13. Gruhl, D., Guha, R., Kumar, R., Novak, J., Tomkins, A.: The predictive power of online chatter. In: Proceedings of the Eleventh ACM Conference on Knowledge Discovery in Data Mining (KDD) (2005)
14. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM (JACM) **46**(5), 604–632 (1999)
15. Landau, L., Segatti, A.: Contemporary migration to South Africa: a regional development issue. World Bank-free PDF (2011)
16. Lerman, K., Jain, P., Ghosh, R., Kang, J.-H., Kumaraguru, P.: Limited attention and centrality in social networks. In: Proceedings of Conference on Social Intelligence and Technology (SOCIETY) (2013)
17. Mao, H., Shuai, X., Ahn, Y.-Y., Bollen, J.: Mobile communications reveal the regional economy in côte divoire. In: Proceedings of the 3rd Conference on the Analysis of Mobile Phone Datasets (NetMob) (2013)
18. Mejova, Y., Srinivasan, P., Boynton, B.: GOP primary season on twitter: popular political sentiment in social media. In: Proceedings of the Sixth ACM Conference on Web Search and Data Mining (WSDM) (2013)
19. Naaman, M., Becker, H., Gravano, L.: Hip and trendy: characterizing emerging trends on twitter. J. Am. Soc. Inform. Sci. Technol. **62**(5), 902–918 (2011)
20. Naveed, N., Gottron, T., Kunegis, J., Alhadi, A.C.: Bad news travels fast: a content-based analysis of interestingness on twitter. In: Proceedings of the Web of Science Conferece (2011)
21. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: linking text sentiment to public opinion time series. In: Proceedings of the 4th AAAI Conference on Weblogs and Social Media (ICWSM) (2010)
22. Quercia, D., Ellis, J., Capra, L., Crowcroft, J.: Tracking gross community happiness from tweets. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW) (2012)

23. Romero, D.M., Galuba, W., Asur, S., Huberman, B.A.: Influence and passivity in social media. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part III. LNCS, vol. 6913, pp. 18–33. Springer, Heidelberg (2011)
24. Ruiz, E.J., Hristidis, V., Castillo, C., Gionis, A., Jaimes, A.: Correlating financial time series with micro-blogging activity. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM) (2012)
25. Scellato, S., Mascolo, C., Musolesi, M., Latora, V.: Distance matters: geo-social metrics for online social networks. In: Proceedings of the 3rd Conference on Online Social Networks (WOSN) (2010)
26. State, W.I., Bogdan, E.Z., et al.: Studying inter-national mobility through ip geolocation. In: Proceedings of the Sixth ACM Conference on Web Search and Data Mining (WSDM) (2013)
27. Taylor, P.J., Ni, P., Derudder, B., Hoyler, M., Huang, J., Lu, F., Pain, K., Witlox, F., Yang, X., Bassens, D., et al.: Measuring the world city network: new developments and results. GaWC Res. Bull. **300** (2009)
28. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: what 140 characters reveal about political sentiment. In: Proceedings of the 4th AAAI Conference on Weblogs and Social Media (ICWSM) (2010)
29. UN. Big Data for Development: A Primer. United Nations, Global Pulse (2013)
30. Ver Steeg, G., Galstyan, A.: Information transfer in social media. In: Proceedings of the 21st ACM Conference on World Wide Web (WWW) (2012)
31. Wellman, B., Haase, A., Witte, J., Hampton, K.: Does the Internet Increase, Decrease, or Supplement Social Capital? Social Networks, Participation, and Community Commitment (2001)
32. Weng, J., Lim, E., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM Conference on Web Search and Data Mining (WSDM) (2010)
33. Weng, L., Flammini, A., Vespignani, A., Menczer, F.: Competition among memes in a world with limited attention. Sci. Rep. **2**(335), 1–8 (2012)